# Application and interpretation of the yo-yo intermittent recovery test to the long-term physical development of girls association football players


| | |
|---|---|
| Journal: | *Science and Medicine in Football* |
| Manuscript ID | RSMF-2018-0119 |
| Manuscript Type: | Original Papers |
| Keywords: | Aerobic fitness, female, youth, LTAD, reliability |



| | |
|---|---|
| Abstract: | We aimed examine yo-yo intermittent recovery test, level 1 (YYIRL1) performance in girls football players. Mixed-linear modelling was used to determine within-season changes in YYIRL1 performance and between- and within- player variation over four years, at four time points (July, September, December and May) in 86 players (474 observations). Twenty-three players, tested over three consecutive years were retained for further analysis. Magnitude-based inferences were used to quantify annual change in performance on a group and individual level. Within-player correlations were used to determine the association between YYIRL1 and maturation. 'Very likely' small (14, ±90% confidence interval 8.1 – 20%) improvements were observed between July and September and 'possibly' trivial (5.5, 0.41 to 11%) differences between September and May. Within-player variation throughout the in-season period, representing the typical error of the estimate was 23%, 22 – 25% and between-player, 38, 33 – 44%. We observed 'most likely' moderate improvements (32, 17 to 49%) over three years that were moderately associated with changes in maturation (r = 0.46, 0.13 to 0.70). A minimum change of ≥44% is required to detect 'likely' improvements in YYIRL1 performance on an individual basis. Girls football players appear responsive to pre-season training and to long-term exposure to systematic training. |

# Application and interpretation of the yo-yo intermittent recovery test to the long-term physical development of girls association football players

We aimed examine yo-yo intermittent recovery test, level 1 (YYIRL1) performance in girls football players. Mixed-linear modelling was used to determine within-season changes in YYIRL1 performance and between- and within- player variation over four years, at four time points (July, September, December and May) in 86 players (474 observations). Twenty-three players, tested over three consecutive years were retained for further analysis. Magnitude-based inferences were used to quantify annual change in performance on a group and individual level. Within-player correlations were used to determine the association between YYIRL1 and maturation. 'Very likely' small (14, ±90% confidence interval 8.1 – 20%) improvements were observed between July and September and 'possibly' trivial (5.5, 0.41 to 11%) differences between September and May. Within-player variation throughout the in-season period, representing the typical error of the estimate was 23%, 22 – 25% and between-player, 38, 33 – 44%. We observed 'most likely' moderate improvements (32, 17 to 49%) over three years that were moderately associated with changes in maturation (r = 0.46, 0.13 to 0.70). A minimum change of ≥44% is required to detect 'likely' improvements in YYIRL1 performance on an individual basis. Girls football players appear responsive to pre-season training and to long-term exposure to systematic training.

Keywords: aerobic fitness, female, youth, LTAD, reliability

## Introduction

The development of the aerobic energy system is important to women's association football, given the requirement to repeat short explosive actions over the course of a match (Datson et al., 2014). Whilst the match demands of girls football are relatively unknown, elite female players cover approximately 10 km within a match and numerous high-intensity bouts (Datson et al., 2014; Datson, Drust, Weston, & Gregson, 2018; Datson et al., 2017). On average, international players cover approximately 608 ± 181 m of high speed running (20 – 25 km.h$^{-1}$) (Datson et al., 2017). High-intensity running bouts are often followed by short periods of recovery (Datson et al., 2018). Recovering from these demanding periods of play, throughout a 90-minute match, is likely to substantially tax the aerobic energy system. Indeed, players achieve mean heart rates of ~86 – 88% of heart rate peak and the amount of high-intensity running performed in a match is associated with performance on the yo-yo intermittent recovery test, level 1 (YYIRL1) (Krustrup, Mohr, Ellingsgaard, & Bangsbo, 2005).

Players at a higher standard of competition are required to perform more high-intensity running than those at lower levels. It has been consistently shown that high-intensity running reduces both between and within halves in female football (Datson et al., 2017; Krustrup, Zebis, Jensen, & Mohr, 2010; Mohr, Krustrup, Andersson, Kirkendal, & Bangsbo, 2008). This highlights the need to both monitor, and develop physical capabilities in female players (Datson et al., 2014). The YYIRL1 is a football-specific field test which highly taxes both the aerobic and anaerobic energy systems and tests players' ability to repeatedly perform aerobic high-intensity work (Bangsbo, Iaia, & Krustrup, 2008; Krustrup et al., 2003). Yo-yo intermittent test performance can differentiate between junior (~17 and 18 years old) and senior female players and between male and female players (Bradley et al., 2012; Mujika, Santisteban, Impellizzeri, & Castagna, 2009). However, the appropriateness of this test for younger (<17 years old) girls football players is yet to be established.

Noticeable lower physical capabilities have been reported in female players in comparison to males (Mohr et al., 2008; Mujika et al., 2009). Senior and junior males have performed 97% and 153% more distance on the YYIRL1 than senior and junior females, respectively (Mujika et al., 2009). However, the gap between the sexes in quantifiable Olympic sports, is less pronounced, approximately 10% (Thibault et al., 2010). Without exposure to systematic training, girls tend to experience lower age-related changes in fitness than boys, with most physical qualities demonstrating a plateaux (Catley & Tomkinson, 2013; Tomkinson et al., 2017). Anatomical and physiological differences, such as post pubertal increases in fat mass, joint laxity and neuromuscular strength may partly explain these observations (Balyi & Hamilton, 2004; Lloyd & Oliver, 2012). Nevertheless, training status is critical in determining physical performance in female players (Krustrup et al., 2005; Mohr et al., 2008; Mujika et al., 2009). Therefore, the development of football-specific fitness is a fundamental component for long-term player development (Wright & Laas, 2016).

It has been shown that children do not lack trainability in aerobic fitness, independent of maturation status when outcomes are normalized for body size (Cunha et al., 2016; McNarry & Jones, 2011). For example, football-specific fitness (yo-yo, level 2 endurance) increases longitudinally in response to sustained and structured training in boys academy football players, compared to age matched controls, and after controlling for maturation status (Wrigley, Drust, Stratton, Atkinson, & Gregson, 2014). Cross-sectional analysis of girls in an FA Regional Talent club show that YYIRL1 performance increases with age groups (under 12, under 14s and under 16s) but these increases are less clear between the older age groups (Emmonds et al., 2018). The longitudinal relationship between improvements in YYIRL1 performance and physical maturation is yet to be established in girls football players who are exposed to systematic training.

The ability to clearly understand if training interventions are making a meaningful difference in physical performance on the pitch is of utmost importance to practitioners and coaches. Accurate interpretation of testing data on an individual level can help tailor training programmes and monitor each player's development. Central to such interpretation of fitness testing data is an understating of reliability (Atkinson & Nevill, 1998) so small but, potentially meaningful changes can be identified on an individual basis. The within-subject variability is likely the most important measure of reliability for the coach or practitioner interested in monitoring performance (Hopkins, 2000). We refer to this statistic as the typical error, which equates to the standard deviation of an individuals' repeated test scores, and can be expressed in raw units or as a percentage, or coefficient-of-variation (Hopkins, 2000). The test-retest typical error of the YYIRL1, in athletes >16 years old, ranges from 4.9 to 13% but appears to be population specific (Schmitz et al., 2018). Indeed, higher typical errors have been reported in girls football players 16% (90 % confidence intervals; 13 to 22%), and in children between 6 to 9 years old using a modified version of the YYIRL1 (19%) (Ahler, Bendiksen, Krustrup, & Wedderkopp, 2011; Wright, Hurst, & Taylor, 2016). Whilst test-retest typical error is useful, reliability refers to the repeatability of a performance on multiple occasions when no systematic improvement is observed (Hopkins, Schabort, & Hawley, 2001; Hurst, Batterham, Weston, & Weston, 2017). To date, such analysis has not been performed on the YYIRL1 within girls football.

We aimed to evaluate the short-term (within-season) and long-term (over four seasons) development of football-specific fitness in girls football players in-line with maturation status. We also wished to estimate the typical error during "in-season" training where the focus is on maintaining, rather than improving football-specific fitness and, use this statistic to interpret changes in performance on an individual level.

## Methods

### *Study design and participant information*

A single-arm quasi-experimental design was used to determine both long-term and short-term changes in YYIRL1 performance. This study was approved by the institution's Research and Governance Ethics Committee (Ethics No. SSSBLREC008). Following medical screening, and obtaining both player and parental consent, we used the YYIRL1 to evaluate football-specific fitness in a total of 86 girls football players from an English FA talent development programme, over four consecutive seasons. The player demographics for each season are shown in Figure 1. Testing was conducted at the start of pre-season (July), and at the start (September), mid-point (December) and end (May) of the English football season. All testing was completed in the same indoor sports hall and at the same time of day. In total 474 observations were made over the four-year period, 111 in July, 122 in September, 122 in December and 119 in May.

INSERT FIGURE 1 HERE

To understand long-term changes in performance we included only players who had completed the YYIRL1 on at least two occasions over a minimum of three consecutive years resulting in 23 players eligible for this analysis (stature, 154 ± 12 cm; mass, 55 ± 17 kg; maturity offset, 0.5 ± 1.4 years from peak-height velocity).

***Procedures***

Mass, standing and sitting stature were measured using a stadiometer and electronic scales (both SECA Medical Measuring Systems, Germany) before each testing session. This data were used with the players' chronological age to predicted biological maturation using the maturity offset, expressed as years from peak-height velocity (Mirwald, Baxter-Jones, Bailey, & Beunen, 2002; Sherar, Mirwald, Baxter-Jones, & Thomis, 2005). A standardised warm-up consisting of light jogging, dynamic mobility and short sprints (including changes of direction) was followed by jumping and sprinting tests which were used for player monitoring purposes (Wright et al., 2016). The YYIRL1 was performed, using the standard procedures (Bangsbo et al., 2008), 10 minutes after the completion of the sprint testing. The test was stopped when a player twice failed to reach the finishing line on time and the highest completed level (e.g. 16.1) was recorded and subsequently converted into distance (meters) before analysis.

***Training and match exposure***

Throughout the in-season time points, players were typically exposed to two 90-minute football training sessions and one 70-minute gym-based strength and conditioning session each week. The latter focused on fundamental movement skill development and neuromuscular training (Wright & Laas, 2016). Approximately 20 fixtures are played throughout a 35-week season lasting between 50-80 minutes. In the pre-season period, repeated-sprint or high-intensity interval training was incorporated with the aim of improving in football specific fitness and preparing the players for the demands of matches.

Direct quantification of training load throughout the four-year period was not possible. Understanding the totality of loading within the chaotic nature of youth team sport is challenging (Phibbs et al., 2017). Indeed, session RPE data load taken over a 17-week period in girls academy football players demonstrated other activities outside of the programme represented 36 ± 10% of the total load (mean 2157 ± 454 Arbitrary Units) and greater within-, than between- player variations in weekly loading were observed (Taylor, Hurst, Best, & Wright, 2015). Furthermore, compliance to training load monitoring was poor (32%) in this study.

***Statistical analysis***

Descriptive statistics are presented as means ± standard deviations. YYIRL1 data were log-transformed for analysis and subsequently back transformed to obtain equivalent percentage values. Mixed linear modelling (SPSS Statistics version 24) was chosen to assess the variability in YYIRL1 performance at each time point (July, September, December and May) and between in-season time points (September, December, May), which were labelled as fixed effects, (September, December, May) with random intercepts to estimate the within- and between- player variation. Changes in year of testing and maturity offset were accounted for as covariates within the model. The within- player variability during in-season time points represented the typical error and 0.2 x the total between- player variability was used to estimate the smallest worthwhile change in YYIRL1 score, statistically.

Identification of a minimal reference for a change in a test is critical to its interpretation. A limitation of using any statically derived values for the smallest worthwhile change is that these are, in essence, only a proxy for a meaningful change in performance on the pitch (Reider, 2015). An alternative method is to select an appropriate anchor (Cook et al., 2014) for example, a 1% change in 20-m sprint time is equivalent to an approximately 20 cm gain in performance on the pitch, which could be the difference between winning and losing the ball in a one-on-one contest (Haugen &

Buchheit, 2016). We also identified the minimal difference in YYIRL1 that we felt was important to football performance. This was four shuttles (160m) which is approximately equivalent to running an extra 90-m at high intensity within a game (Krustrup et al., 2005).

Long-term changes in YYIRL1 performance between the players' 1st, 2nd and 3rd year within the talent development programme were analysed using a customised Microsoft Excel spread sheet (Hopkins, 2006). The relationship between changes in maturation status and YYIRL1 performance was assessed by within-subject correlation (Bland & Altman, 1995) through a general linear model (SPSS Statistics version 24) with the uncertainty of the estimate expressed as 90% confidence intervals. Changes on an individual level were analysed by inputting the players' YYIRL1 score, the typical error, degrees of freedom and our minimum reference value for change into a separate customised spread sheet (Hopkins, 2017). Players were identified as "responders" if they demonstrated a *likely* (>75%) chance of a positive or negative change. The number and proportion (%) of responders are presented with the 95% confidence intervals, calculated using the Wilson method (Newcombe, 1998).

Magnitude based inferences (Hopkins, Marshall, Batterham, & Hanin, 2009) were applied to all differences between time points. The uncertainty of the estimate was calculated from the disposition of the 90% confidence interval to the smallest worthwhile change. Verbal descriptors were assigned using the following scales: 0.5–5 % very unlikely; 5–25 % unlikely; 25–75 % possibly; 75–95 % likely; 95–99.5 % very likely; > 99.5 % most likely. The effect was deemed unclear if the confidence interval overlapped the smallest positive or negative change by ≥ 5 % (Batterham & Hopkins, 2006). Given the chance of inferential error increases with multiple observations, the uncertainty of the estimate was also evaluated conservatively via disposition of 99%

confidence interval (Hopkins, 2007). Inferences remaining 'likely' are indicated with bold text in Figure 2. The magnitude of the differences were evaluated through standardised differences in the means using the following thresholds: < 0.2 trivial; < 0.6 small; < 1.2 moderate; < 2 large; < 4 very large; ≥ 4 extremely large (Hopkins et al., 2009).

**Results**

Mean YYIRL1 performance (m), within- and between- player variability for each time period, and changes between each time period are presented in Figure 2.

INSERT FIGURE 2

Players performed 733 ± 240m in July, 844 ± 283m in September, 824 ± 267m in December and 895 ± 246m in May. Within-player variability throughout the in-season time points (typical error) was 23% (90% confidence interval, 23%, 22 to 25%) and between-player was 38, 33 – 44%, resulting in a smallest worthwhile change of 67m.

Players performed 817 ± 301m in year one of the talent development programme, 968 ± 369m in year two and 1075 ± 358m in year three. We observed a moderate within-player correlation between YYIRL1 score and maturation over this period (r = 0.46, 0.13 to 0.70) for which the individual correlations are displaced in Figure 3.

INSERT FIGURE 3

The changes in YYIRL1 performance between each year are presented on a group and individual basis in Figure 4.

INSERT FIGURE 4

Out of the 23 players we observed likely improvements beyond 67m in 13 (57%, 95% confidence interval, 37 to 74%), and conversely, likely substantial decrements in two (8.7, 2.4 to 27%) (Figure 3A). Nine (39, 22 to 59%) of those players were likely to have improved beyond 160m (Figure 3B). The individual change in YYIRL1 performance, with 90% confidence limits and magnitude-based inferences are presented in figure 5.

INSERT FIGURE 5

**Discussion**

We present a comprehensive evaluation of variability in YYIRL1 performance in girls football players, demonstrating changes in football-specific fitness throughout a season and over consecutive seasons. Our findings support those of a recent meta-analysis indicating male football players perform worst on the YYIRL1 at initial pre-season testing, but improve throughout the pre-season period (Bangsbo et al., 2008; Schmitz et al., 2018). Indeed, our data suggest that most of the within-season improvements were observed in the pre-season period. We found moderate improvements over pre-season, similar to those reported previously in male rugby union players (Mclaren, Smith, Bartlett, Spears, & Weston, 2018) and in girls football players (Wright et al., 2016), supporting the notion that girls football players respond to periods of targeted training (Wright et al., 2016). Furthermore, girls have demonstrated a plateaux in cardiorespiratory fitness post-puberty when not exposed to structured training (Catley & Tomkinson, 2013; Tomkinson et al., 2017). In contrast we observed ‘likely’, small improvements between both year 1 and year 2 and between year 2 and year 3. Cumulatively, these improvements were ‘likely’ to be substantially greater than the four shuttles (160 m) we estimated to reflect an important change with regards to match physical performance.

YYIRL1 performance ranged within season from 733 ± 240 to 895 ± 246 m, similar values have been reported in other girls talent development programmes, where performance ranged from 635 ± 241 in the under 12's to 959 ± 299 in under 16's age group (Emmonds et al., 2018). We observed 'most likely', moderate improvements in players who had been within the programme for three consecutive seasons, supporting previous research (Emmonds et al., 2018). Furthermore, these improvements were only moderately associated with maturation. Far stronger within-player correlations have been shown previously between maturation and sprint related physical qualities (Wright & Atkinson, 2017). This suggests that other factors are also likely to influence the development of football specific fitness in this cohort. Differences in training loads represent one such factor that could influence changes in YYIRL1 performance. In particular, central load, measured using differential ratings of perceived exertion has been associated with improvements in YYIRL1 performance in professional rugby union players (Mclaren et al., 2018). Unfortunately, it is difficult to capture the totality of training load in youth players given the chaotic nature of loading in these athletes (Phibbs et al., 2017; Taylor et al., 2015). Unpublished data from our practice suggest perceived central exertion is substantially higher in fitness sessions, typical of the pre-season period, than in football training and matches or strength and conditioning, typical of the in-season period.

***Interpretation of individual change in YYIRL1 performance***

A noticeable limitation of our study was the lack of an appropriate comparator group. It can be difficult to recruit appropriate age-matched controlled participants for applied research, particularly over the full duration of the study (Atkinson & Batterham, 2017; Wright & Atkinson, 2017). An important function of a control group is to account for 'noise' in the data. When analysing individual changes in performance, without a control group, it is important to account for random within-player variations in the measure (Atkinson & Batterham, 2017). Our individual analyses accounted for the within-player variation over the in-season period as the typical error for YYIRL1 performance in this group. No systematic improvements were expected over this period and indeed any differences on a group level between September and May were trivial.

We chose to interpret our data using the magnitude-based inference approach, which some statisticians suggest may increase Type 1 errors (Sainani, 2018; Welsh & Knight, 2015). These criticisms have been defended in detail (Hopkins & Batterham, 2016; Hopkins & Batterham, 2018). Given the fundamental limitations to null-hypothesis testing (Page, 2014) and the difficulties in choosing an appropriate prior distribution for YYIRL1 performance in girls football players to inform a fully Bayesian analysis, magnitude-based inferences, by choosing a dispersed uniform prior, provides an appropriate approach in this instance. Moreover, it is the most relevant method to analysing an individual's performance (Buchheit, 2018). We also mitigate against the rising chance of inferential error associated with multiple observations by also evaluating the MBI's most conservatively, based upon the disposition of the 99% confidence interval. Finally, it is important to note that findings from any single-arm study should be interpreted with caution.

Our study is at the applied end of the basic-applied research continuum (Atkinson & Nevill, 2001). A strength of this type of research is that it better represents the delivery environment within a girls talent development centre but consequently has less control of extraneous variables. For example, we scheduled testing around habitual training practices. The data generated from such an approach may be more applicable to coaches and practitioners than those generated in a highly controlled setting (Enright et al., 2018). Our typical error was higher (23%) when compared to those reported in athletes over 16 years of age (4.9 to 13%) (Schmitz et al., 2018) and compared to test-retest data in a similar cohort (16%) (Wright et al., 2016). A similar observation has been previously seen in men's football where the test-retest coefficient-of-variation was 5% however, the typical error during the season, where no mean group changes were observed was 15% (Krustrup et al., 2003). This suggests that test-retest typical error, separated by a short period of time may underestimate the typical variation in YYIRL1 performance.

We identify a greater proportion of players who improved YYIRL1 performance than those who didn't (Figure 4). However, the high typical error reported made the identification of responders and non-responders difficult given the width of the 90% confident intervals (Figure 5). Some players with moderate improvements did not reach our threshold for a 'likely' responder (> 75% chance of improvement). Indeed, an increase of 32% is required before a practitioner can detect a 'likely' change in YYIRL1 performance. In such instances there is still a 12% chance of a negative response. When applying non-clinical magnitude-based inferences, a positive effect would normally be deemed "clear" when the chance of a negative effect is less than 5% or visa-versa (Hopkins et al., 2009), in this case a change of ≥56%. However, with measures where the typical error is greater than the smallest worthwhile change this may be too conservative for individual player analysis (Hopkins, 2017). Thus, a coach or practitioner has to make a judgement as to the level of uncertainty they deem appropriate to identify player improvements.

A second key judgement a practitioner must make in interpreting individual data is the choice of the smallest worthwhile change. To detect a 'likely' change beyond 4 shuttles (160 m) a change of ≥44% is required, resulting in fewer players being identified as 'likely' responders (figure 4B). Irrespectively, these data demonstrate the sensitivity of the YYIRL1 to detect true individual changes in performance when used in girls football players.

### *Conclusions and practical applications*

Girls football players appear responsive to both targeted short-term, pre-season training and long-term exposure to systematic training as part of a talented development center. Long-term improvements were moderately associated with changes in maturation status suggesting other factors, such as training load, effect the long-term development of football-specific fitness. These data support the notion that dedicated fitness training enhances YYIRL1 performance (Wright et al., 2016) and supports previous recommendations that aerobic development should be strategically planned (Emmonds et al., 2018; Wright & Laas, 2016). However, practitioners need to consider the appropriateness of their training prescriptions based upon a player's holistic athletic development. Thus the development of fundamental movement skill or neuromuscular strength, speed and co-ordination also need to be considered (Lloyd & Oliver, 2012; Wright & Laas, 2016).

Time efficient strategies to improve football-specific fitness in-season would be beneficial to enhance the development of physical qualities in girls football players and close the observed gap in physical performance between male and female players (Mujika et al., 2009). For example, repeated-sprint training (e.g. 3 to 4 sets of 7 x 30-m sprints with 20 seconds recovery) is a time efficient method for enhancing both YYIRL1 and sprint performance (Taylor, Mclaren, & Weston, 2016) which can be easily interspersed with technical or tactical training.

Finally, we demonstrated a high typical error for the YYIRL1, which was greater than the smallest worthwhile change and limits the sensitivity of the test when monitoring individual players. A change of >44% is required to detect likely changes in YYIRL1 performance beyond four shuttles. Practitioners may wish to consider other tests, such as the multi-stage fitness test or the 30:15 intermittent fitness test to monitor their players.

Figure 1. Characteristic of the players included within the study over each of the four seasons.

Figure 2: Mean YYIRL1 performance for each time point. Between- and within-player variation is expressed as a standard deviation. Differences between time points are presented with magnitude-based inference, percentage difference (90% confidence interval). Qualitative inferences are indicated as 'possibly' (P), 'likely' (L), 'very likely' (VL) and 'most likely' (ML). Bold text represents a change that is 'likely' based upon the disposition of the 99% confidence interval in relations to the smallest-worthwhile change.

Figure 3: Individual within-player correlations between YYIRL1 performance and maturity offset.

Figure 4: Group and individual changes in YYIRL1 performance over three years with magnitude-based inference and percentage difference (90% confidence interval). Red and green markers indicate individual responders. The smallest worthwhile change was either 0.2 between-subject standard deviations (A) or four shuttles (B).

Figure 5: Individual changes in YYIRL1 performance over three years with 90% confidence intervals. Qualitative inferences are indicated as 'unclear' (?), 'likely' (L), 'very likely' (VL) and 'most likely' (ML). The smallest worthwhile change was either 0.2 between-subject standard deviations (A) or four shuttles (B).

## References

Ahler T, Bendiksen M, Krustrup P, Wedderkopp N. 2011. Aerobic fitness testing in 6- to 9-year-old children: reliability and validity of a modified Yo–Yo IR1 test and the Andersen test. *Eur J Appl Physiol. 112*(3):871–876. http://doi.org/10.1007/s00421-011-2039-4

Atkinson G, Batterham AM. 2017. The impact of random individual differences in weight change on the measurable objectives of lifestyle weight management services. *Sports Med*. 47(9):1683–1688. http://doi.org/10.1007/s40279-017-0683-5

Atkinson G, Nevill AM. 2001. Selected issues in the design and analysis of sport performance research. *J Sports Sci.* 19(10):811–827. http://doi.org/10.1080/026404101317015447

Atkinson G, Nevill AM. 1998. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med.* 26(4):217–238. http://doi.org/10.2165/00007256-199826040-00002

Balyi I, Hamilton A. 2004. Long-term athlete development: Trainability in childhood and adolescence. Windows of opportunity. Optimal trainability. Victoria. National Coaching Institute British Columbia & Advanced Training and Performance Ltd.

Bangsbo J, Iaia FM, Krustrup P. 2008. The Yo-Yo intermittent recovery test: a useful tool for evaluation of physical performance in intermittent sports. *Sports Med.* 38(1):37–51.

Batterham AM, Hopkins WG. 2006. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*. 1(1):50–57.

Bland JM, Altman DG. 1995. Statistics notes: Calculating correlation coefficients with repeated observations: Part 1—correlation within subjects. *Br Med J.* 310(6977):446. http://doi.org/10.1136/bmj.310.6977.446

Bradley PS, Bendiksen M, Dellal A, Mohr M, Wilkie A, Datson N, et al. 2012. The application of the Yo-Yo intermittent endurance level 2 test to elite female soccer populations. *Scand J Med Sci Sports*. 24(1):43–54. http://doi.org/10.1111/j.1600-0838.2012.01483.x

Buchheit M. 2018. A battle worth fighting: a comment on 'The vindication of magnitude-based inference'. *Sport Perform Sci Reviews.* 31(1):1-2.

Catley MJ, Tomkinson GR. 2013. Normative health-related fitness values for children: analysis of 85347 test results on 9-17-year-old Australians since 1985. *Br J Sports Med.*47(2):98–108. http://doi.org/10.1136/bjsports-2011-090218

Cook J, Hislop J, Adewuyi T, Harrild K, Altman D, Ramsay C, et al. 2014. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technol Asses*. 18(28):1–202. http://doi.org/10.3310/hta18280

Cunha GDS, Vaz MA, Geremia JM, Leites GT, Baptista RR, Lopes AL, Reischak-Oliveira Á. 2016. Maturity status does not exert effects on aerobic fitness in soccer players after appropriate normalization for body size*. Pediatr Exerc Sci*. 28(3):456–465. http://doi.org/10.1123/pes.2015-0133

Datson N, Drust B, Weston M, Gregson W. 2018. Repeated high-speed running in elite female soccer players during international competition. *Sci Med Football*. Epub. http://doi.org/10.1080/24733938.2018.1508880

Datson N, Drust B, Weston M, Jarman IH, Lisboa PJ, Gregson W. 2017. Match physical performance of elite female soccer players during international competition. J *Strength Cond Res.* 31(9):2379–2387. http://doi.org/10.1519/JSC.0000000000001575

Datson, N., Hulton, A., Andersson, H., Lewis, T., Weston, M., Drust, B., & Gregson, W. 2014. Applied Physiology of Female Soccer: An Update. *Sports Med.*44(9):1225–1240. http://doi.org/10.1007/s40279-014-0199-1

Emmonds, S., Till, K., Redgrave, J., Murray, E., Turner, L., Robinson, C., & Jones, B. 2018. Influence of age on the anthropometric and performance characteristics of high-level youth female soccer players. *Int J Sports Sci Coa.* Epub. http://doi.org/10.1177/1747954118757437

Enright K, Morton J, Iga J, Lothian D, Roberts S, Drust B. 2018. Reliability of "in-season" fitness assessments in youth elite soccer players: a working model for practitioners and coaches. *Sci Med Football*. 2018;2(3):177–183. http://doi.org/10.1080/24733938.2017.1411603

Haugen T, Buchheit M. 2016. Sprint running performance monitoring: Methodological and practical considerations. *Sports Med.* 46(5):641–456. http://doi.org/10.1007/s40279-015-0446-0

Hopkins WG, Marshall SW, Batterham AM, Hanin J. 2009. Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc.* 41(1):3–13. http://doi.org/10.1007/s40279-016-0517-x

Hopkins WG, Schabort EJ, Hawley JA. 2001. Reliability of power in physical performance tests. *Sports Med.* 31(3):211–234.

Hopkins WG. 2007. A Spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p value. *Sportscience* [Internet]. 11:16–20. Available from: sportsci.org/2007/wghinf.htm

Hopkins WG. 2017. A Spreadsheet for monitoring an individual's changes and trend. Sportscience [Internet]. *Sportscience*. 21:5–9. Available from: sportsci.org/2017/wghtrend.htm

Hopkins WG. 2000. Measures of reliability in sports medicine and science. *Sports Med*. 30(1):1–15.

Hopkins WG. 2006. Spreadsheets for the analysis of controlled trials, with adjustment for a subject characteristic [Internet]. *Sportscience* 10:46–50. Available from: sportsci.org/2006/wghtests.htm

Hurst C, Batterham AM, Weston KL, Weston M. 2017. Short- and long-term reliability of leg extensor power measurement in middle-aged and older adults. *J Sports Sci.* 36(9):970–977. http://doi.org/10.1080/02640414.2017.1346820

Krustrup P, Mohr M, Amstrup T, Rysgaard T, Johansen J, Steensberg A, et al. 2002. The yo-yo intermittent recovery test: physiological response, reliability, and validity. *Med Sci Sports Exerc.* 35(4):697–705. http://doi.org/10.1249/01.MSS.0000058441.94520.32

Krustrup P, Mohr M, Ellingsgaard H, Bangsbo J. 2005. Physical demands during an elite female soccer game: importance of training status. *Med Sci Sports Exerc.* 37(7):1242–1248. http://doi.org/10.1249/01.mss.0000170062.73981.94

Krustrup P, Zebis M, Jensen JM, Mohr M. 2010. Game-induced fatigue patterns in elite female soccer. *J Strength Cond Res.* 24(2):437–441. http://doi.org/10.1519/JSC.0b013e3181c09b79

Lloyd RS, Oliver JL. 2012. The Youth Physical Development Model: A new approach to long-term athletic development. *Strength Cond J.* 34(3):61–72. http://doi.org/10.1519/SSC.0b013e31825760ea

Mclaren SJ, Smith A, Bartlett JD, Spears IR, Weston M. 2018. Differential training loads and individual fitness responses to pre-season in professional rugby union players. *J Sports Sci.* 36(21):2438–4246. http://doi.org/10.1080/02640414.2018.1461449

McNarry M, Jones A. 2011. The influence of training status on the aerobic and anaerobic responses to exercise in children: A review. *Eur J Sport Sci.* 2011;14 Suppl 1:S57–S68. http://doi.org/10.1080/17461391.2011.643316

Mirwald RL, Baxter-Jones ADG, Bailey DA, Beunen GP. 2002. An assessment of maturity from anthropometric measurements. *Med Sci Sports Exerc.* 34(4):689–694.

Mohr M, Krustrup P, Andersson H, Kirkendal D, Bangsbo J. 2008. Match activities of elite women soccer players at different performance levels. *J Strength Cond Res.* 22(2):341–349. http://doi.org/10.1519/JSC.0b013e318165fef6

Mujika DI, Santisteban J, Impellizzeri FM, Castagna C. 2009. Fitness determinants of success in men's and women's football. *J Sports Sci.* 27(2):107–114. http://doi.org/10.1080/02640410802428071

Newcombe RG. 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med.* 17(8):873–890.

Page, P. 2014. Beyond statistical significance: clinical interpretation of rehabilitation research literature. *Int J Sports Physical Therapy*. 9(5): 726 – 736.

Phibbs PJ, Jones B, Roe G, Read D, Darrall-Jones J, Weakley J, et al. 2017. Organised chaos in late specialisation team sports: Weekly training loads of elite adolescent rugby union players. *J Strength Cond Res.* Epub. http://doi.org/10.1519/JSC.0000000000001965

Reider B. 2015. Good, or Just Better? *Am J Sports Med*. 43(8):1841–1843. http://doi.org/10.1177/0363546515595612

Schmitz B, Pfeifer C, Kreitz K, Borowski M, Faldum A, Brand S-M. 2018. The Yo-Yo Intermittent Tests: A systematic review and structured compendium of test results. *Front Physiol*. 9:776–716. http://doi.org/10.3389/fphys.2018.00870

Sherar LB, Mirwald RL, Baxter-Jones ADG, Thomis M. 2005. Prediction of adult height using maturity-based cumulative height velocity curves. *J Pediatr.* 147(4):508–514. http://doi.org/10.1016/j.jpeds.2005.04.041

Taylor JM, Hurst C, Best R, Wright MD. 2015. Contribution of planned and unplanned training to overall load in elite youth female football players. *World Congress on Science and Soccer*, Copenhagen, Denmark; 23rd–24th May.

Taylor JM, Mclaren SJ, Weston M. 2016. Two weeks of repeated-sprint training in soccer: To turn or not to turn? *Int J Sports Physiol Perform*. 11(8):998–1004. http://doi.org/10.1123/ijspp.2015-0608

Thibault V, Guillaume M, Berthelot G, Helou NE, Schaal K, Quinquis L, et al. 2010. Women and men in sport performance: The gender gap has not evolved since 1983. *J Sports Sci Med.* 9(2):214–223.

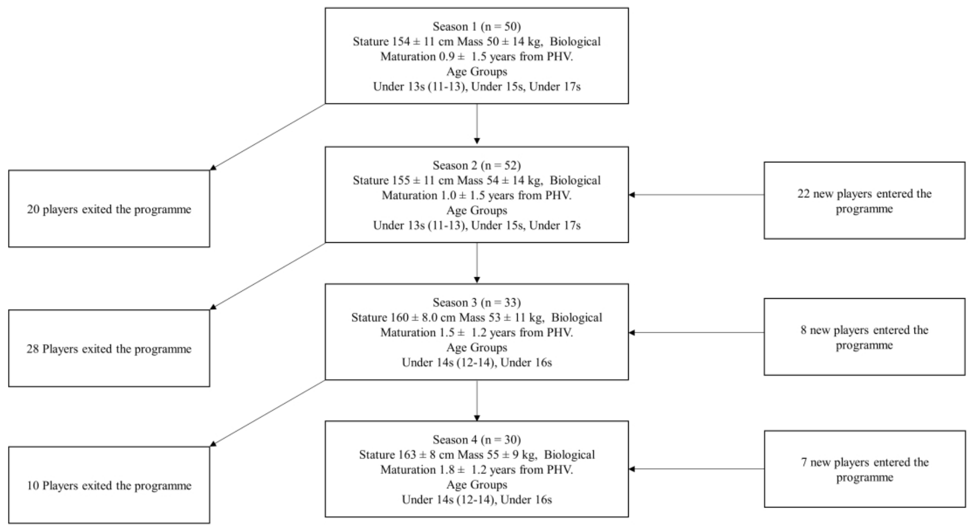
Tomkinson GR, Carver KD, Atkinson F, Daniell ND, Lewis LK, Fitzgerald J S, et al. 2017. European normative values for physical fitness in children and adolescents aged 9-17 years: results from 2 779 165 Eurofit performances representing 30 countries. *Br J Sports Med*. Epub.

Wright MD, & Atkinson G. 2017. Changes in sprint-related outcomes during a period of systematic training in a girls' soccer academy. *J Strength Cond Res.* 2017. Epub. http://doi.org/10.1136/bjsports-2017-098253

Wright MD, Hurst C, Taylor JM. 2016. Contrasting effects of a mixed-methods high-intensity interval training intervention in girl football players. *J Sports Sci.* 34(19):1808-1815. http://doi.org/10.1080/02640414.2016.1139163

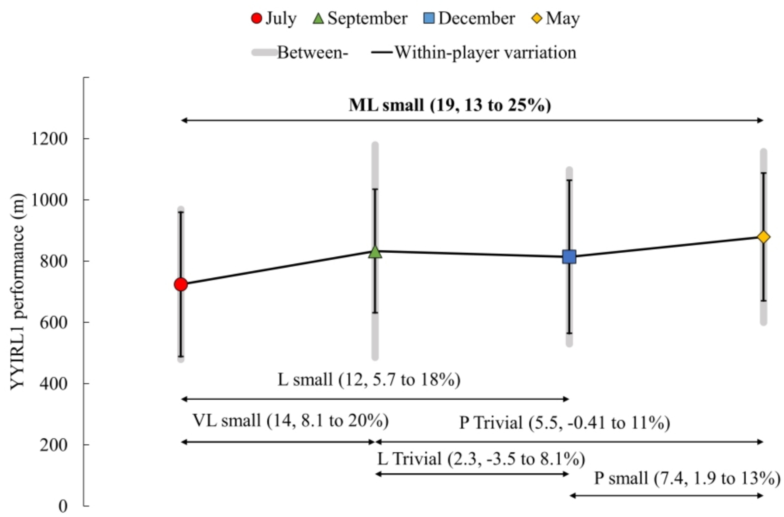
Wright MD, Laas M-M. 2016. Strength Training and Metabolic Conditioning for Female Youth and Adolescent Soccer Players. *Strength Cond J.* 38(2):96–104. http://doi.org/10.1519/SSC.0000000000000212

Wrigley R, Drust B, Stratton G, Atkinson G, Gregson W. 2014. Long-term soccer-specific training enhances the rate of physical development of academy soccer players independent of maturation status. *Int J Sports Med.* 35(13):1090–1094. http://doi.org/10.1055/s-0034-1375616



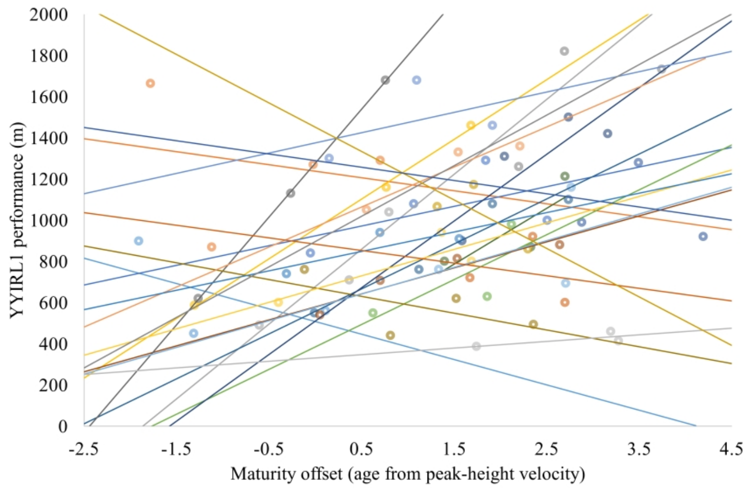

Characteristic of the players included within the study over each of the four seasons.

338x190mm (300 x 300 DPI)



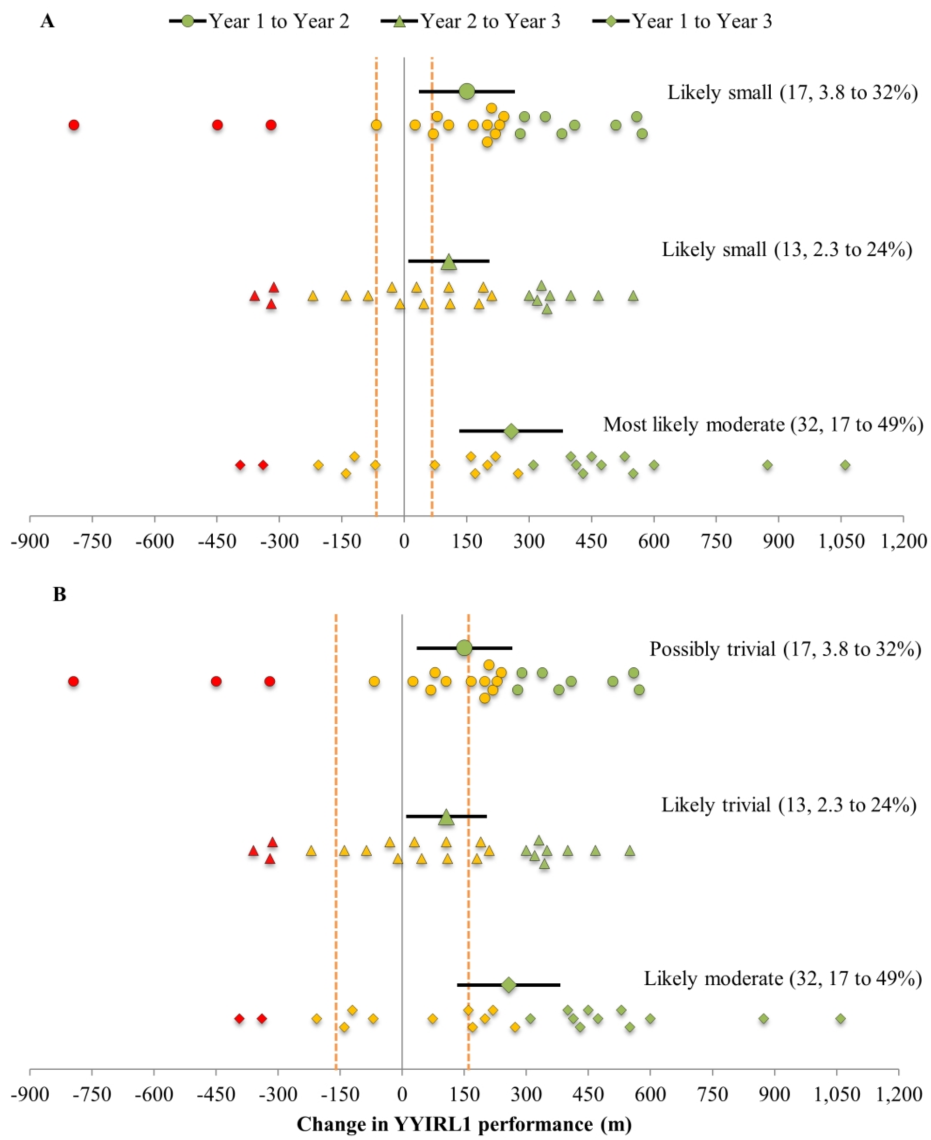

Mean YYIRL1 performance for each time point. Between- and within-player variation is expressed as a standard deviation. Differences between time points are presented with magnitude-based inference, percentage difference (90% confidence interval). Qualitative inferences are indicated as 'possibly' (P), 'likely' (L), 'very likely' (VL) and 'most likely' (ML). Bold text represents a change that is 'likely' based upon the disposition of the 99% confidence interval in relations to the smallest-worthwhile change.

338x190mm (300 x 300 DPI)



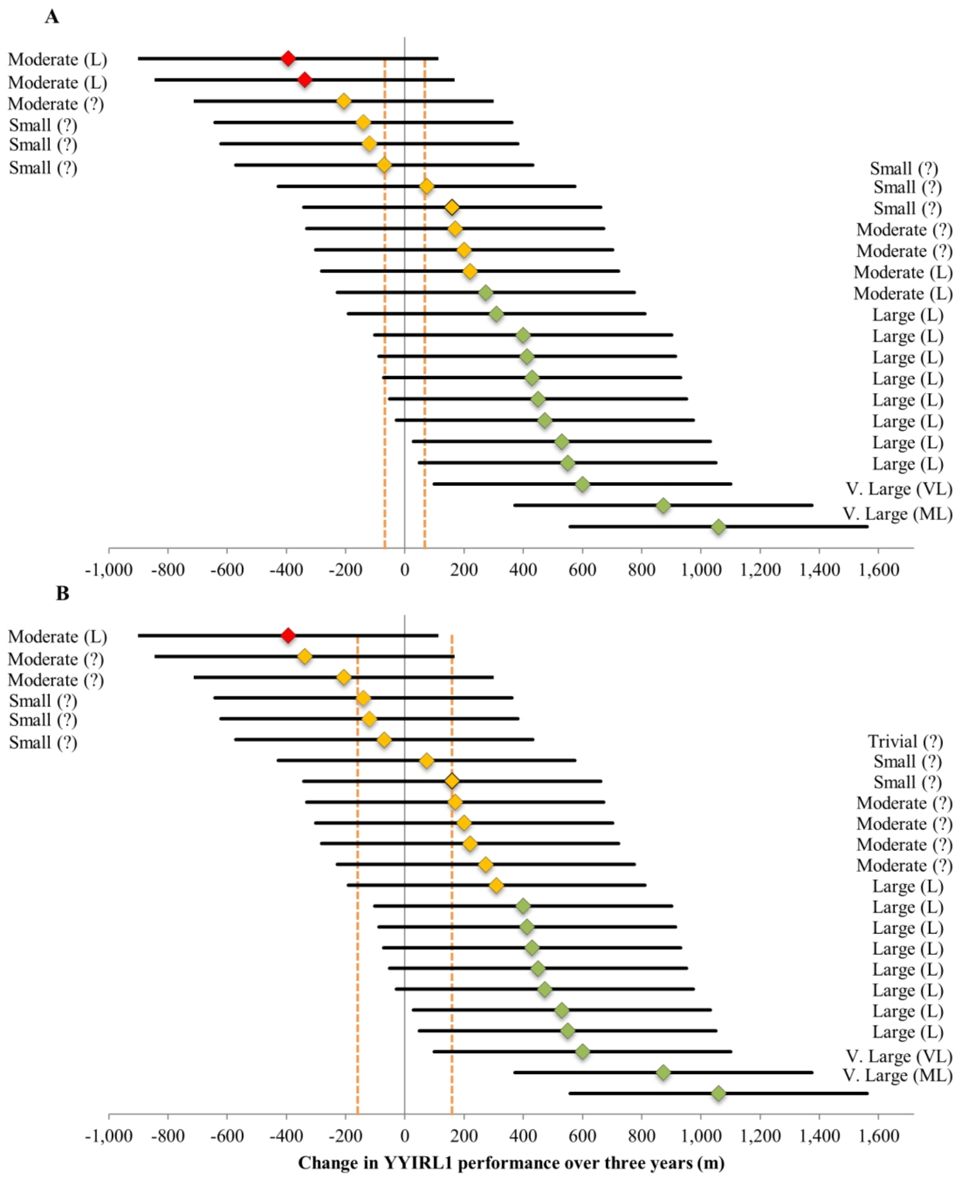

Individual within-player correlations between YYIRL1 performance and maturity offset.

338x190mm (300 x 300 DPI)




Group and individual changes in YYIRL1 performance over three years with magnitude-based inference and percentage difference (90% confidence interval). Red and green markers indicate individual responders. The smallest worthwhile change was either 0.2 between-subject standard deviations (A) or four shuttles (B).

190x238mm (300 x 300 DPI)




Individual changes in YYIRL1 performance over three years with 90% confidence intervals. Qualitative inferences are indicated as 'unclear' (?), 'likely' (L), 'very likely' (VL) and 'most likely' (ML). The smallest worthwhile change was either 0.2 between-subject standard deviations (A) or four shuttles (B).

190x238mm (300 x 300 DPI)